基因表达系列分析技术的新进展

李 靖[1*] 陈宇光[1] 孔祥银[2]

[1](上海大学生命科学学院,上海 200436)

[2](中国科学院上海生物工程中心,上海 200233)

摘 要 作为新近建立的研究基因表达的有效工具,基因表达系列分析(SAGE)技术能同时对大量的转录物进行定性和定量分析。它不仅可以显示低丰度的转录物,提供基因组表达的完整信息,而且可以通过不同状态下基因表达图谱的比较,深入了解基因表达的时空性和有序性,从而寻找和发现新基因。本文介绍了 SAGE 的工作原理和方法,并着重对其最新的应用与研究进展进行了综述。

关键词 基因表达,基因表达系列分析(SAGE)

中图分类号 Q753 **文献标识码** C **文章编号** 1000-3061(2001)06-0613-04

基因表达水平的分析对于研究及了解生物体特性和基因功能起着至关重要的作用。目前,用于研究基因表达水平的方法主要有 EST(Expressed sequence tag)序列鉴定、差异显示(Differential display,DD)、微阵列杂交(Microarray)、消减杂交(Substraction hybridization,SH)和基因表达序列分析(Serial analysis of gene expression,SAGE)。其中,SAGE 法是一项高通量且快捷有效的基因表达研究技术,可用于研究任何一种由细胞转录变化引起的生物现象,而无须对基因性质和生物系统预先有所了解。任何一个具备 PCR 和手动测序仪器的普通实验室都能使用这项技术,而且操作中锚定酶与标签酶的不同组合使之更具灵活性。因此,SAGE 法[1] 自 1995 年问世以来引起了研究人员的广泛关注。

1 基因表达系列分析技术的原理和操作

SAGE 技术主要基于两条原则:首先,来自转录物内特定位置的 9 ~ 10 bp 短核苷酸序列(SAGE 标签)所含信息足以代表其相应的转录物;其次,SAGE 标签经随机连接、扩增并集中在一个克隆中测序,标签重复出现的次数代表该转录物的拷贝数。具体操作方法是:(1)提取 mRNA,以生物素化的寡核苷酸(Oligo dT)为引物反转录合成 cDNA 后,用锚定酶(Anchoring enzyme,AE)消化。通过链霉亲和素蛋白珠(Streptavidin beads)收集 cDNA 的 3′端部分,将其作为该转录物的独特信息。(2)将所得 cDNA 等分为两部分,分别与含标签酶(Tagging enzyme,TE)位点的接头 A 或接头 B 连接。这种标签酶是典型的 II 类限制酶,能在其不对称识别位点上游的 20 bp 处切割 DNA 双链,产生平末端。(3)用标签酶酶切产生连有接头的短 cDNA 片段(约 9 ~ 10 bp),混合两个 cDNA 池,待短cDNA片段钝端相连构成双标签(Ditag)后进行 PCR 扩增。(4)用锚定酶切割扩增产物,分离双标签并克隆、测序。一般每个克隆的标签数处于 10 ~ 50 之间。(5)对标签数据进行 SAGE 软件分析,并与 Genbank、ESTdatabases 等数据库比较,获得转录物丰度的数量信息及新的表达基因。

2 基因表达系列分析技术的应用

SAGE 技术以其使用灵活,便于掌握,能进行全基因组表达水平分析,适用于比较不同状态下基因表达情况和有助于发现新基因的独特属性获得了广泛应用。

2.1 全面获取生物基因的表达信息

SAGE 区别于差异显示、消减杂交等其它技术的主要特点是可用于寻找那些较低丰度的转录物,最大限度地收集基因组的基因表达信息。这使之成为从总体上全面研究基因表达、构建基因表达图谱的首选策略。

1995 年,Velculescu 等[1] 首先用 *Nls*III 作锚定酶、*Bsm*FI 作标签酶,对人胰腺组织基因表达进行了分析,得到了人胰腺组织的基因表达图谱。1997 年,Velculescu 等[2] 又报道了由 SAGE 方法对不同生长期酵母转录物的分析数据构建的染色体表达图谱。结合基因表达信息与基因组图谱绘制的染色体表达图谱,使基因表达与物理结构融合起来,更利于基因表达模式的研究。2001 年,Huib Caron 等人[3] 在已知人类基因染色体定位的基础上,用 SAGE 方法构建了特定细胞类群中表达基因及其转录水平的转录组(Transcriptome)图并发现了众多低丰度的表达基因和在特定染色体区域内的高表达基因簇。目前,该图谱中包括了 12 种组织的 245 万个 SAGE 转录标签,其中含 160 000 个成神经细胞瘤标签。与绝

收稿日期:2001-04-26,修回日期:2001-06-20。

* 通讯作者。 Tel:86-21-66743286; E-mail:leean@eastday.com

大多数为高表达基因的 EST 序列相比，SAGE 标签提供了更多低丰度的表达信息。此外，人们还先后创建了人卵母细胞、单核细胞来源的树状细胞、小神经胶质细胞、胎儿成纤维细胞、CD15 + 骨髓先祖细胞等多种细胞及人类甲状腺组织、脑组织、肾组织、肝组织等多种组织和水稻小苗等各种不同的表达图谱，对广泛表达的普通基因和在少数细胞及组织中表达的特定基因作了详细描述。

2.2 定量比较不同状态下的基因表达

SAGE 法同样可用于在不同环境，不同生理状态及不同生长阶段的细胞和组织表达图谱构建。对不同状态下基因表达水平的定量或定性比较，特别是对疾病组织与正常组织的比较发展迅速。2000 年，Hough 等[4] 用 SAGE 技术系统分析了不同卵巢癌细胞和组织的转录特点。385 000 个 SAGE 标签被分析，代表了 56 000 多个转录物。与正常组织相比，一些基因的表达发生了明显变化，其中发生表达上调的有网格蛋白-3，4、粘蛋白-1 等表面及分泌蛋白，此外有关脂类稳定的脱辅基脂蛋白 E 和 J 的表达也出现了高度上调。实验结果经随后的免疫组化杂交证实。目前，使用这一策略已发现了许多在癌症中呈现上调表达的基因。这些基因，尤其是正常组织中缺乏的基因，很可能成为有用的诊断和预测指标或潜在的治疗位点。而且，对于那些通过肿瘤抑制基因或癌基因来调控转录水平的基因鉴别将为瘤形成分子机制的了解提供有益的帮助。同时，SAGE 技术也应用于其它疾病组织中，如对粥样硬化[5]、HIV 感染[6]、唐氏综合症[7] 的研究已提供了不少有关发病机理的线索。

另外，Welle 等人[8,9] 还将 SAGE 法用于基因表达变化对肌肉生理功能的潜在作用研究及人、鼠肌肉中与衰老相关的基因表达变化比较，证实了衰老时肌肉功能的下降与基因转录的变化密切相关且因物种而异。Robert-Nicoud 等[10] 也用 SAGE 技术分析了醛固酮和抗利尿激素对鼠肾皮质细胞的作用，揭示了与之相关的细胞钠、水吸收的调节机制。在含 Kringe 结构的新跨膜蛋白 Kremen 的分子克隆与特征研究中，Nakamura 等[11] 同样以 SAGE 方法取得了巨大的成功。总之，无论使用何种材料、研究何种状态，针对何种方向，只要是基于基因表达水平的分析，SAGE 技术就大有用武之地，这正是 SAGE 的最大优势所在。

2.3 寻找新基因

在以上所述的构建基因表达图谱和对基因表达的定量比较过程中都可能发现新基因。一方面，通过所得 SAGE 标签与已知基因的对比分析，若发现标签没有同源序列，那么这个标签就很可能与新基因相对应。此时，用 SAGE 标签作探针筛选 cDNA 文库，就可能钓出新基因。另一方面，对不同状态下表达图谱的比较会发现基因表达上的差异。这些差异往往是由于细胞或组织所处的状态所致，差异基因很可能就是与之相关的新基因。如 Inadera 等[12] 比较了用 17-β-雌激素处理前后的 MCF-7 乳腺癌细胞，发现两类细胞表达图谱的绝大部分相同，但有 4 个基因表达量增加，其中 3 个是已知的刺激素诱导基因——组织蛋白酶 D、pS2 和高流动性组 1 蛋白。另一个是最近在哺乳动物上皮细胞系 C57MG 中报道的由癌基因 Wnt-1 变化而来的上调基因 WISP-2，它直接调节刺激素受体，是一个雌激素反应型的新基因。Xu 等[13] 利用同样的方法研究了雄激素调控基因并在前列腺中发现了位于染色体 20q13 区域上高度表达的新基因——PMEPA1。

通过 SAGE 方法寻找新基因只是新基因发现过程中的第一步，对每个预计新基因的分析和了解最终还须在综合的整体条件下加以深化和印证。

3 基因表达系列分析技术的研究进展

SAGE 方法虽具有显著的优势，但作为一种全新的技术还存在着不足。为使 SAGE 更加实用可靠，应用更广，人们对它进行了系统的研究，提出了不少行之有效的改进方法和补充方案。

3.1 SAGE 操作优化

对实验材料的数量要求是 SAGE 技术的一大问题，使之局限于细胞系等大量起始材料的表达分析，而对活组织检查等少量或微量的生物样本很难见效。同时，由于特殊细胞亚型的表达改变可能会被同时存在于组织中的其它细胞的表达稀释，对高度异化的细胞群落组成的复杂组织进行分析也很困难。针对这些问题，Datson[14] 提出了 microSAGE 解决方案。microSAGE 法将 SAGE 中从 RNA 分离到标签释放的所有步骤简化为"单管"程序，在一个试管中实现。这不仅将原始 RNA 用量减少至 1/500 ~ 1/5000，也提高了效率，减少了纯化步骤。另外，该法还采用有限的 PCR 循环次数获得足量双标签。实验证明，microSAGE 法可用于放大大脑中的特异表达区域，获得特异性表达图谱，尤其适合与微解剖技术联合使用对少量、微量材料和异源组织进行分析。Peters 等人[15] 也为此进行了 SAGE 的改进，提出了 SAGE-Lite 法，将实验材料的初始用量进一步减少到 0.1 μg，同时也使 SAGE 标签与长 cDNA 更易快速分离。2000 年，Ye S Q 等[16] 提出了 mini-SAGE 法。尽管在材料用量方面，该法不及 microSAGE 和 SAGE-Lite 少，但却有其独到之处：(1)取消附加的 PCR 扩增，彻底避免了 SAGE 中可能由 PCR 扩增引入的定量误差。(2)降低连接时接头加入量，尽量减少接头对 SAGE 双标签扩增的干扰，从而增加了双标签的产量。(3)在 DNA 酚抽提后，使用锁相胶提高 DNA 的纯度和收率。他们用此法成功建立了两个成纤维细胞的 SAGE 标签库，对其中一个库的 3838 个标签的初步分析证实了一个典型的成纤维细胞基因表达图谱。

另外，虽然 SAGE 技术能最大限度地全面收集生物组织的基因表达信息，但也不能完全保证涵盖所有的低丰度 mRNA。而且，标签体的连接可能会因为接头的干扰造成克隆中标签体过少及克隆序列末端不能高效连入载体。1998 年 Powell[17] 对此进行了改进。他用生物素化的 PCR 引物扩增双标签，随后与链霉亲和素蛋白珠结合除去多余接头，从而使双标签的数量及克隆产物的信息量剧增。而 Kenzel-

mann 等[18]的改进方案则是在电泳前增加加热步骤来富集并获得较长的连接体以降低 SAGE 成本。此外,Angelastro 等[19]还发现,在 *Nla*III 酶消化双标签前加入单纯化步骤,可在不减少 DNA 数量和不影响随后连接的条件下,酶切产生更多的双标签。

3.2 SAGE 数据库与分析软件

对 SAGE 法产生的庞大数据的综合分析有赖于数据库及系统软件的帮助。随着 SAGE 技术的发展,此类数据库和软件层出不穷。

USAGE[20]是一个基于网络的包含全套 SAGE 分析工具的软件包。它可用于分析和比较各种 SAGE 数据,也能为新的 SAGE 实验提供参考帮助。USAGE 不仅提供多用户环境,使用户间能共享数据,同时也提供与相关数据库的链接界面,便于用户储存和分析结果。用户可登陆 http://www.cmbi.kun.nl/usage/免费使用 USAGE,体会它的强大功能与便捷服务。另外,与之类似的综合性工具软件还有 eSAGE[21]等。

ExProView[22]软件通过二维点阵列的形式将转录数据分类,形成虚拟芯片。芯片上的每个点均代表一个在表达基因解剖数据库或 UniGene 中定义的已知基因。通过对阵列点的选择,用户可从本地数据库或 WWW 连接中获取表达基因的更多信息。该软件不仅在描述水平上提供了 mRNA 表达的整体模型,对功能表达模型的开发做出指导,而且可以直接提供每个基因的详尽信息。获取 ExProView 相关信息可至 http://www.biochem.kth.se/exproview。

POWER_SAGE[23]是一种在两个样品间检测标签表达频率差异的统计学软件。该程序能以样本大小和标签频率的不同组合进行 SAGE 的仿真研究,并可对每种组合的有效性加以判断,是设计 SAGE 实验的有效工具。

作为公用基因表达资源,SAGEmap[24]是癌症基因组解剖计划(Cancer Genome Anatomy Project,CGAP)的一部分,主要用于放置、获取和分析人类基因表达数据。该数据库提供 WWW 和 FTP 两个位点,它们的网址分别为 http://www.ncbi.nlm.nih.gov./sage 和 ftp://ncbi.nlm.nih.gov/pub/sage。此外,通过登陆 http://www.ncbi.nim.nih.gov/SAGE > SAGE,使用者还可对任何两个已知数据库中的转录物进行对比分析。

尽管如上所述,使用数据库和分析软件对 SAGE 标签进行分析有众多好处,但必须认识到,由于 EST 序列错误和其 5′和 3′方向倒置的原因,软件分析可能出错,对于标签的分配最终仍须手工确认。

3.3 SAGE 与其它技术的联用

为了保证结果的可靠性,在实验中 SAGE 技术往往与基因芯片或 Northern 杂交联合使用。此外,针对 SAGE 技术的局限性,一些配合使用的新策略也应运而生。

对于 SAGE 标签序列在基因识别应用上的两大问题:(1)标签的长度不足,很难进行深入的基因描述;(2)一些特定的标签序列与数据库中的大量序列匹配,使得在这些匹配序列中确定特定组织内与 SAGE 标签相对应的正确序列十分困难,Chen 等[25]创立了利用 SAGE 标签产生较长 cDNA 片段来识别基因的新方法(GLGI)。该技术的主要特征是:以包含 SAGE 标签的序列为有义引物,以锚定寡核苷酸(Oligo dT)为无义引物,用 *Pfu* DNA 聚合酶进行 PCR 扩增。通过该途径,SAGE 标签序列可以立即转变为相应 cDNA 上从 SAGE 标签到 3′末端区域间的较长 cDNA 片段。GLGI 技术为 SAGE 的更广泛使用进行了有效的补充,两者联用不仅能确定人类或其它真核生物基因组中表达基因的 3′端边界,构建基因表达的完整目录,而且也可用于对基因组测序中以生物信息学工具预测的外显子的真实性确认。无独有偶,Anke van den Berg 等[26]为较长 cDNA 片段的获得提供了另一思路,他们建立了用 mRNA RT-PCR 法产生较长 cDNA 片段的 SAGE 标签快速分析法(RAST-PCR)。

另外,针对基因组中大多数基因的低水平表达,Wang 和 Powell[27]在 SAGE 方法的基础上引入差减克隆等基因鉴定技术,形成了综合性基因鉴定程序(Integrated procedure for gene identification,IPGI)。该程序中增加的差减反应减少了 SAGE 中的冗余拷贝,极大地降低了测序分析的工作量;增加的抑制性 PCR 选择性扩增富集了稀有拷贝,确保了信息的完整性。同时,此方法还强调了与已有 EST 信息文库的紧密联系,使发现基因的能力大大增强。

4 结　语

随着人类基因组计划和模式生物基因组测序工作的逐渐完成,在大量序列信息的支持下,不断完善和发展的 SAGE 技术必将充分显示其巨大的应用潜能,为包括环境基因组、药物基因组在内的功能基因组研究提供新手段,在全面综合了解基因表达与生命活动关系和后基因组计划中发挥更大的作用。

REFERENCES(参考文献)

[1] Velculescu V E, Zhang L, Vogelstein B *et al*. Serial analysis of gene expression. *Science*, 1995, **270**(5235): 484 ~ 487

[2] Velculescu V E, Zhang L, Zhou W *et al*. Characterization of the yeast transcriptome. Cell, 1997, **88**(2): 243 ~ 251

[3] Huib C, Barbera van Schaik, Merlijn van der Mee *et al*. The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science*, 2001, **291**(5507): 1289 ~ 1292

[4] Hough C D, Sherman-Baust C A, Pizer E S *et al*. Large-scale serial analysis of gene expression reveals genes differentially expressed in ovarian cancer. *Cancer Res*, 2000, **60**(22): 6281 ~ 6287

[5] De Waard V, van den Berg B M, Veken J *et al*. Serial analysis of gene expression to assess the endothelial cell response to an atherogenic stimulus. *Gene*, 1999, **226**(1): 1 ~ 8

[6] Ryo A, Suzuki Y, Ichiyama K *et al*. Serial analysis of gene expression in HIV-1-infected T cell lines. *FEBS* Letters, 1999, **462**(1 ~ 2): 182 ~ 186

[7] Chrast R, Scott H S, Papasavvas M P *et al*. The mouse brain transcriptome by SAGE: differences in gene expression between P30 brains of the partial trisomy 16 mouse model of Down syndrome (Ts65Dn) and normals. *Genome Res*, 2000, **10**(12): 2006 ~ 2021

[8] Welle S, Bhatt K, Thornton C A. High-abundance mRNA in human muscle: comparison between young and old. *Journal of applied physiology*, 2000, **89**(1): 297 ~ 304

[9] Welle S, Brooks A, Thornton C A. Senescence-related changes in gene expression in muscle: similarities and differences between mice and man. *Physiol Genomics*, 2001, **5**(2): 67 ~ 73

[10] Robert-Nicoud M, Flahaut M, Elalouf J M *et al*. Transcriptome of a mouse kidney cortical collecting duct cell line: effects of a aldosterone and vasopressin. *Proc Natl Acad Sci USA*, 2001, **98**(5): 2712 ~ 2716

[11] Nakamura T, Aoki S, Kitajima K *et al*. Molecular cloning and characterization of Kremen, A novel kringle-containing transmembrane protein. *Biochim Biophys Acta*, 2001, **19**, 1518(1 ~ 2): 63 ~ 72

[12] Inadera H, Hashimoto S, Dongl H Y *et al*. WISP-2 as a novel estrogen-responsive gene in human breast cancer cells. *Biochemical and Biophysical Research Communication*, 2000, **275**(1): 108 ~ 114

[13] Xu L L, Shanmugam N, Sesterhenn I A *et al*. A novel androgen-regulated gene, PMEPA1, located on chromosome 20113 exhibit high level expression in protstate. *Genomics*, 2000, **66**(3): 257 ~ 263

[14] Datson N A, van der Perk-de Jong J, van den Berg M P *et al*. MicroSAGE: a modified procedure for serial analysis of gene expression in limited amounts of tissue. *Nucleic Acids Res*, 1999, **27**(5): 1300 ~ 1307

[15] Peters D G, Kassam A B, Yonas H *et al*. Comprehensive transcript analysis in small quantities of mRNA by SAGE-lite. *Nucleic Acids Res*, 1999, **27**(24): 39

[16] Ye S Q, Zhang L Q, Zheng F *et al*. MiniSAGE: gene expression profiling using serial analysis of gene expression from 1 microg total RNA. *Anal Biochem*, 2000, **287**(1): 144 ~ 152

[17] Powell J. Enhanced concatemer cloning-a modification to the SAGE (Serial Analysis of Gene Expression) technique. *Nucleic Acids Res*, 1998, **26**(14): 3445 ~ 3446

[18] Kenzelmann M, Muhlemann K. Substantially enhanced cloning efficiency of SAGE (Serial Analysis of Gene Expression) by adding a heating step to the original protocol. *Nucleic Acids Res*, 1999, **27**(3): 917 ~ 918

[19] Angelastro J M, Klimaschewshi L P, Vitolo O V. Improved NlaⅢ digestion of PAGE-purified 102bp ditages by addition of a single purification step in both the SAGE and microSAGE protocols. *Nucleic Acids Res*, 2000, **28**(12): 62

[20] Van Kampen A H, van Schaik B D, Pauws E *et al*. USAGE: a web-based approach towards the analysis of SAGE data. *Bioinformatics*, 2000, **16**(10): 899 ~ 905

[21] Margulies E H, Innis J W. eSAGE: managing and analysis data generated with serial analysis of gene expression. *Bioinformatics*, 2000, **16**(7): 650 ~ 651

[22] Larsson M, Stahl S, Uhlen M *et al*. Expression Profile Viewer (ExProView): A software tool for transcriptome analysis. *Genomics*, 2000, **63**(3): 341 ~ 353

[23] Man M, Wang X, Wang Y. POWER-SAGE: comparing statistical tests for SAGE experiment. *Bioinformatics* 2000, **16**(11): 953 ~ 959

[24] Lash A E, Tolstoshev C M, Wagner L *et al*. SAGEmap: a public gene expression resource. *Genome Res*, 2000, **10**(7): 1051 ~ 1060

[25] Chen J J, Janet D R, Wang S M. Generation of longer cDNA fragments from serial analysis of gene expression tags for gene identification. *Proc Natl Acad Sci USA*, 2000, **97**(1): 349 ~ 353

[26] Anke van den Berg, Judith van der Lei, Sibrand P. Serial analysis of gene expression: rapid RT-PCR analysis of unknown SAGE tags. *Nucleic Acids Research*, 1999, **27**(17): 17e

[27] Wang S M, Janet D R. A strategy for genome-wide gene analysis: Integrated procedure for gene identification. *Proc Natl Acad Sci USA*, 1998, **95**(20): 11900 ~ 11910

New Progress of Serial Analysis of Gene Expression

LI Jing[1]* CHEN Yu-Guang[1] KONG Xiang-Yin[2]

[1] (*School of Life Sciences, Shanghai University, Shanghai* 200436, *China*)

[2] (*Shanghai Research Center of Biotechnology, Chinese Academy of Sciences, Shanghai* 200223, *China*)

Abstract As an efficient tool that has been developed recently, serial analysis of gene expression (SAGE) allows the qualitative and quantitative analysis of a large number of transcripts. It can define the transcripts at relatively low levels and characterize the genomic expression near completion. In addition, it provides insights into the timely and orderly expression of genes by comparing the profiles constructed for a pair of cells that are kept at different conditions, thus identifying a set of novel genes. In this review, the newest progress of SAGE's application and research is mentioned in details with its original method and principle outlined.

Key words gene expression, serial analysis of gene expression(SAGE)

Received: April 26, 2001

* Corresponding author. Tel: 86-21-66743286; E-mail: leean@eastday.com